Eur J Psychol Educ (2017) 32:133–156
DOI 10.1007/s10212-015-0282-5

Teachers' reasons for using peer assessment: positive experience predicts use

Ernesto Panadero[1] · Gavin T. L. Brown[2]

Received: 25 August 2015 / Revised: 21 December 2015 / Accepted: 23 December 2015 /
Published online: 18 January 2016

Abstract Peer assessment (PA) is one of the central principles of formative assessment and assessment for learning (AfL) fields. There is ample empirical evidence as to the benefits for students' learning when AfL principles are implemented. However, teachers play a critical role in mediating the implementation of intended policies. Hence, their experiences, beliefs, and attitudes towards PA are important factors in determining whether the policy is actually carried out. A survey of over 1500 primary, secondary, and higher education teachers in Spain elicited their beliefs and values around PA as well as other aspects of formative assessment; only 751 teachers provided complete responses to all PA items. Teachers reported occasional use of PA in their classrooms but with positive experience of it. The vast majority did not use anonymous forms of PA and half of the teachers considered the students were accurate when assessing peers. Confirmatory factor analysis and structural equation modeling were used to examine relationships of attitudes and beliefs to self-reported frequency of using of PA. The self-reported frequency of using PA was strongly predicted by teacher experience of PA which included positive reasons for using PA, rather than negative obstacles for avoiding, prior use, and beliefs that students should participate in assessment, and willingness to include PA in grading.

Keywords Peer assessment · Assessment for learning · Formative assessment · Teachers' conceptions · Assessment conceptions

✉ Ernesto Panadero
ernesto.panadero@uam.es

[1] Departamento de Psicología Evolutiva y de la Educación, Universidad Autónoma de Madrid, Madrid, Spain

[2] Faculty of Education and Social Work, The University of Auckland, Auckland, New Zealand

Introduction

Among the practices of formative assessment (FA) and assessment for learning (AfL), peer assessment (PA) occupies a central role (Black and Wiliam 1998; Nicol and Macfarlane-Dick 2006). PA is a process through which a student considers the characteristics of a peer's performance or work according to appropriate criteria and standards so as to judge the quality and characteristics of the work (Topping 2013). PA is central to AfL and FA because it involves students in assessment, generates feedback that might be useful to the evaluated peer, and also gives the peer assessor insights as to how their own work might be improved (i.e., an indirect self-assessment) (Dochy et al. 1999; Reinholz 2015). Additionally, PA helps students to become more self-regulating and use more advanced learning strategies when performing a task (Topping 2003). While, there is research about peer assessment effects on learning (e.g., Dochy et al. 1999) and its reliability and validity (e.g., Topping 2003), there is less research on classroom use of PA. Specifically, we have a poor understanding of the reasons teachers have for using or not using PA; the goal of this study. This is important because assessment activities are mediated by the teaching professional responsible for classroom activities.

Understanding the beliefs, values, and attitudes of teachers who are responsible for implementing assessment practices in general and who, specifically, have been challenged to introduce assessment practices aimed at contributing to the development of student self-regulated learning strategies and processes seems an important step in educational psychology. Beliefs and attitudes, combined with awareness of social norms and perceived levels of control, are important contributors to intentions and behavior (Ajzen 2005). Teacher beliefs guide teachers in understanding educational policies, deciding what is important, and determining what should be done (Fives and Buehl 2012). Specifically, teacher beliefs about assessment practices have been shown to vary according to level of schooling in which teachers work and their cultural norms, but generally there is little evidence that teachers provide opportunities for students to engage in peer assessment (Barnes, Fives, and Dacey 2015). Thus, understanding the reasons and experiences teachers use to implement a specific assessment practice will provide insights to both policy and professional development processes that ought to respond to the actual beliefs teachers have. Examination of these reasons in a new cultural context will also shed light on the degree to which previous studies can be generalized.

Teachers' peer assessment use and beliefs

It is important to understand the thinking, values, beliefs, and reasons teachers have in order to understand their use of PA that is recommended by curriculum and policy. While more is known of how students experience PA, less is known about teachers' perceptions of PA. This is important since, notwithstanding student involvement in PA, it is the teacher who initiates and implements PA in classroom settings.

The most usual finding is that teachers value PA as a learning activity. For example, Bryant and Carless (2010) found that Hong Kong teachers considered PA helpful in students learn to write; PA constituted a "wider skill" that empowered students to monitor their own performance independent of the teacher. Harris and Brown (2013) found among three New Zealand teachers that they were aware of the potential of PA to enhance students' learning and self-regulated learning. Likewise, Noonan and Duncan (2005) found that "some" Canadian teachers preferred PA because it facilitated group work and/or activities. Lynch and Golen (1992) explored PA use by 78 USA business university teachers and found two interesting aspects. First, 54 % perceived PA as effective in improving the students' writing skills or their attitude towards writing. Second, PA was perceived as having a number of strengths, including (a) bringing out the best in students, (b) improving the quality of reports, (c) adding variety and pace to teaching, (d) exposing students to different people's judgment, (e) improving writing and grades, (f) helping students gain respect for others' opinions, (g) helping students learn how to give feedback, and (h) helping students gain confidence in their ability to judge another's writing. Hence, it would appear reasonably safe to conclude that teachers are generally aware that through PA students can develop greater understanding of criteria and standards and use that to improve their own work.

However, there is some much lesser awareness among teachers that PA would impact relationships between the teacher and the students and the relationships among their own students. Only one of the three teachers observed in New Zealand was aware of potential risks in the use of PA which was addressed by active intervention (Harris and Brown 2013). Noonan and Duncan (2005) found that, out of 110 Canadian high school teachers, 49 % reported little use and just 27 % reported some use of PA and self-assessment, a result partly explained by their concern that high school students lacked the "maturity to be truthful and/or objective in peer and self-assessment" (p. 5). Despite the positive perceptions of the teachers in Lynch and Golen's (1992) study, PA was perceived as having a number of weaknesses, including (a) student reluctance to criticize the work of their friends, (b) inflation of scores when the work was not anonymous, (c) student resistance to the grading process, (d) need to exercise caution and diplomacy, and (e) disproportional benefits to the good, perceptive students rather than weaker students.

The concerns that teachers express about accuracy in PA (e.g., Lynch and Golen 1992; Noonan and Duncan 2005) refers to the relationship between the teacher's judgment or score and the peer's score for the same piece of work. A lack of inter-rater consensus has implications for the validity of an assessment practice such as PA (for a detailed discussion, see Panadero et al. 2013). Concerns about the validity of student PA are founded in the idea that students, as novices and learners, may not be sufficiently competent in a field to make an accurate estimation of another student's work quality. Empirical work has generally established that PA can be a reliable source of information about students' performance (Falchikov and Goldfinch 2000; Topping 2003). However, while teachers express concern about student accuracy in PA, this belief may not be a deciding factor when teachers choose to implement PA in the classroom.

Topping (1998) indicated privacy is an important aspect of PA, in that disclosing one's identity as opposed to being anonymous, seems to matter to students. Some studies have found students feeling more positive when anonymity was assured (Vanderhoven et al. 2015), while others have found both advantages and disadvantages under conditions of anonymity (Yu and Wu 2011). Nevertheless, the role that anonymity plays for teachers is more uncertain; Harris and Brown (2013) found one teacher insisting that students not be anonymous in their peer marking and commentary on student writing to ensure proper social interrelations and sensitivity towards others. Consequently, it is important to explore whether teachers are keeping PA processes anonymous in the classroom.

In sum, teachers are aware of a number of positive and negative aspects of PA, but how these variables relate to implementation of PA in the classroom is not known. Additionally, our knowledge about teacher perceptions of PA is based on a very small sample of studies, with relatively small numbers of participants, and sometimes quite narrowly specialized teachers (e.g., university business teachers). Therefore, there is need for a larger study of teacher perceptions of PA practices at all educational levels, which is the main aim of this paper.

Peer assessment explored effects

However, because perceptions may not be accurate, it is useful to briefly review what has been empirically established about the nature and function of PA. Peer assessment has received attention in a series of reviews (Falchikov and Goldfinch 2000; Panadero 2016; Topping 1998, 2013; van Gennip et al. 2009; van Zundert et al. 2010). Topping (1998) identified 17 elements of PA that matter to understanding what and how PA is being done. Although some PA research at the primary and secondary education level exists, the bulk of the empirical evidence comes from higher education (van Gennip et al. 2009; van Zundert et al. 2010). A more recent review (Panadero 2016) has found one PA study with primary students, a second with primary, intermediate, and secondary students, a third with girls aged 11 to 18 in a comprehensive rural school, three studies with vocational school students, and one with secondary education students. These numbers point out the enormous difference in the knowledge base regarding how PA is implemented in primary and secondary education, a matter addressed in this study.

Implementing PA in the classroom benefits learning and performance, problem-solving skills, metacognition, and self-regulated learning (Hwang et al. 2014; Kim and Ryu 2013; Spandorfer et al. 2014; Panadero et al. 2016) and can even have advantages over teacher assessment (Falchikov 1995; Topping 2003; Van Gennip 2012). Nevertheless, good PA requires structure and guidance, such as rubrics that seem to improve the quality of PA (Panadero et al. 2013).

Nonetheless, PA is not without some disadvantages. Peers may not easily accept responsibility for assessing peers (Bryant and Carless 2010; Gao 2009; Harris and Brown 2013); Peterson and Irving 2008; Topping 1998); students may perceive PA as unfair (Carvalho 2012); and peers may not feel safe exposing their work to peers or receiving an evaluation from a peer (van Gennip et al. 2009; van Gennip et al. 2010). This latter phenomenon has been identified as privacy in PA (Topping 1998) referring to whether the PA is conducted anonymously, confidentially, or publicly. Differential effects in PA have been found when it is anonymous or public (Vanderhoven et al. 2015), although definitive conclusions about the effect of anonymity as beneficial or harmful for learning cannot be made (Panadero 2016).

Finally, the official weight of PA refers to whether the PA score contributes to a student's final or overall grade (Topping 1998). This matters since contributing towards final grade moves PA from a purely learning exercise to a summative accountability evaluation. When PA counts, it is more important that student scores are trustworthy and it is more likely that interpersonal relationships (e.g., friendship marking) will contaminate the scores. Hence, making PA count may be counterproductive to learning and constitute a reason to resist rather than implement PA.

Spanish context and its assessment practices

Spain is an interesting context in which to study the implementation of PA because of the history of educational legislation. In 1990, a major restructuring of Spanish education via the Ley Orgánica de Ordenación General del Sistema Educativo (LOGSE; Organic Law for the Education System Organization) took place. This law promoted new methodological approaches towards formative assessment purposes in the compulsory K-12 school system (Remesal 2007, 2011) so as to focus attention on evaluating competencies and not just examination results. This was extended in the next reform act (i.e., 2002 Ley Orgánica de Calidad de la Educación (LOCE; Organic Law for Educational Quality)). More recently, two new reform acts (Ley Orgánica de Educación (LOE; Organic Law for Education) 2006 and Ley Orgánica para la Mejora de la Calidad Educativa (LOMCE; Organic Law for the Improvement for Educational Quality) 2013) have moved assessment back to a more summative approach, via external evaluation. Hence, it is interesting to explore the impact that the more formative assessment reform acts (LOGSE and LOCE) might have had an impact on teacher self-reported use of PA in primary and secondary education. It is unlikely that the more recent legislation (LOE and LOMCE) will have had much of an impact on data collected in 2012.

The Spanish context may be enlightening for other jurisdictions since the legal policy around assessment and evaluation seems to have been impacted by two major policy reforms. First, it was impacted by formative assessment and assessment for learning reforms which are fairly widespread globally due to widely publicized claims of effectiveness for learning gains (Black and Wiliam 1998). Second, it shows evidence of more conservative policies to evaluate schools through external evaluation mechanisms (Lingard and Lewis 2016). A further advantage of the Spanish case is that it is not from the English-speaking world which has dominated published research on both assessment for learning reform and school accountability assessment. These characteristics, then, may be instructive for other non-English speaking jurisdictions seeking to weigh up the merits of the two policy reform processes.

The Spanish higher education context does not have clear guidelines about what type of assessment should be implemented, with each individual teacher or department taking decisions about assessment practices. Additionally, Spanish university teachers do not have compulsory specific training on pedagogical aspects, and a previous study found these teachers implementing traditional approaches to assessment (i.e., exams and written work) (Ion and Cano 2011). Additionally, previous research exploring Spanish teachers' student self-assessment practices, another crucial aspect of FA, found significant differences between primary, secondary, and university teachers (Panadero et al. 2014).

Aim and research questions

The present study explores Spanish primary, secondary, and higher education teachers' self-reported implementation of PA and their reasons for its use. A goal of the study was to determine whether the concerns raised in the literature exist and affect how frequently teachers report using PA. The research questions are:

RQ1—What do teachers think about the various aspects of PA?
This RQ was further divided into four more specific RQs for clarity of presentation:
RQ1a—Do Spanish teachers report using PA, with what frequency, what is their experience, and what is the preferred privacy format?
RQ1b—Do they consider PA accurate? Why?
RQ1c—Would teachers let a percentage of their course grade depend on PA score? Which percentage? Why?
RQ1d—What are the main advantages and disadvantages of PA?
RQ2—How do these beliefs influence self-reported PA implementation?

It is hypothesized that (1) teacher perceptions of PA would influence their use of PA, (2) perceptions of the learning benefits of PA would increase its use, (3) perceived difficulties in PA (e.g., student immaturity, interpersonal problems, lack of expertise) would decrease use of PA, (4) use of anonymous forms of PA would increase PA use, and (5) giving official weight for PA would reduce PA use.

RQ3—What differences exist between educational levels?

It is hypothesized that (1) PA would be reported as happening more frequently among K-12 teachers than university teachers in accordance with assessment for learning policies, and (2) the structural relations influencing PA frequency of use would be the same across all levels.

Method

A non-experimental, anonymous, self-report survey, consisting of both open response and fixed-format rating items, was self-administered by a national sample of Spanish teachers.

Participants

A total of 1312 primary education institutions, 814 secondary education institutions, and 7 public universities were contacted. Educational institutions were contacted via phone ($n = 677$) or via e-mail ($n = 1456$). Two selection criteria for institutions were utilized: first, the goal was to maximize the number of institutions because it was likely few teachers were employed at each centre; second, the goal was to maximize the geographic representativeness of the K-12 schooling system. To achieve this, the Education Department's websites for all Spanish regions were searched for files containing lists of institutions. In most regions, the lists were accessible, but the level of information varied; some lists included centre name, phone, and email addresses, others just name and phone number, and others just a list of centre names.

A total of 1286 teachers participated; of these, 441 were primary, 690 were secondary, and 155 were university teachers. Responses from the whole sample were used for RQ1. However, only 751 (39.9 % primary teachers, 52.7 % secondary education, and 7.3 % higher education teachers) indicated they had ever used peer assessment and only their responses were used for RQ2. Finally, for RQ3 both sample sizes were used.

In terms of demographic information, the 1286 teachers had an average of 18 years of experience (SD = 10.07, range 0–43). Two thirds ($N = 861$) were female, 235 were male (21 %), with the remainder not reporting their gender ($N = 160$, 12 %). In terms of teaching subjects, 441 (34 %) teachers taught arts and humanities; 249 (19 %) mostly primary teachers reported teaching multiple subjects at the same time; formal sciences, math, and chemistry 241 (19 %); social sciences 222 (17 %); technical disciplines (e.g. architecture, engineering and computer science) 92 (7 %); health sciences 38 (3 %); and 3 had missing data. In terms of the teacher training, over 80 % had taken general pedagogy courses, about 70 % of primary and secondary teachers had taken assessment courses, whereas only 40 % of university teachers had. In contrast, only 15–20 % of primary and secondary teachers had taken formative assessment courses, in contrast to 35 % of university teachers. This latter result probably arises because the university from which the majority of HE participants belonged has implemented since 2010 a training program on formative assessment. Additionally, in Table 1 the distribution by Spanish regions can be seen, with all 17 regions included in this study. Excluding the missing cases, the

Table 1 Frequency and percentage of participants by Spanish regions

Region	Frequency	Percent	Population	Ranking by population	% population
Andalucía	210	16.3	8,401,567	1	18.12
Madrid	179	13.9	6,377,364	3	13.74
Comunidad Valenciana	125	9.7	4,939,550	4	10.65
Murcia	91	7.1	1,463,249	10	3.15
Cataluña	82	6.4	7,504,008	2	16.17
Asturias	70	5.4	1,049,754	13	2.26
Castilla-Leon	62	4.8	2,478,376	6	5.34
Extremadura	59	4.6	1,091,591	12	2.35
País Vasco	48	3.7	2,164,311	7	4.66
Canarias	40	3.1	2,128,647	8	4.59
Aragón	33	2.6	1,325,385	11	2.86
Baleares	31	2.4	1,124,744	14	2.42
Navarra	29	2.3	636,638	15	1.37
Castilla-la Mancha	14	1.1	2,078,611	9	4.48
Cantabria	11	.9	588,656	16	1.27
Galicia	6	.5	2,734,915	5	5.89
La Rioja	5	.4	313,615	17	0.68
Missing	191	14.9	–	–	–

Note. $N = 1286$

correlation between percent of sample by province and percent of national population by region is $r = .81$ suggesting that there is a reasonable overlap ($R^2 = .65$) between region size and actual obtained sample. The chi-square difference test between the percent of people in sample to percent of nation in each region had $p = .05$ which supports the claim that the distribution of the sample is not statistically different to the distribution of the population.

Given an average of 33.61 teachers per K-12 centre,[1] the maximum possible respondents, if all institutions had chosen to participate, would be 71,453. Assuming the sample is completely random, the current study sample of 1131 K-12 teachers has a margin of error of just 2.89 % in estimating the population values for the various variables of interest. Unfortunately, the small sample of higher education instructors relative to the estimated population of instructors in seven universities (i.e., 155/10,824) has a margin of error of 7.84 %; this means that observed values reported for the higher education sample is much less characteristic of the population than that of the K-12 sample. Finally, because our sample was dependent on the centre leaders' commitment to distribute the survey, and because participation was voluntary, the sample is considered a convenience one.

Instrument

A self-report survey instrument including a total of 75 questions concerning teachers' assessment practices and conceptions was administered. The questionnaire was organized around

[1] As published by the Spanish Educational Department in 2014 (https://www.mecd.gob.es/servicios-al-ciudadano-mecd/estadisticas/educacion/indicadores-publicaciones-sintesis/cifras-educacion-espana/2014.html).

nine blocks: (a) demographic information, (b) testing, (c) scoring, (d) feedback, (e) self-assessment, (f) peer assessment, (g) tools for assessment, (h) emotions related to assessment, and (i) institutional assessment culture. In this study we only present data related to peer assessment in light of participant demographic information. As can be seen in the instrument (Appendix), ten different questions formed the PA survey block. At the beginning of the PA block, the definition, taken from Topping (1998, p. 250), of PA was presented: "Peer assessment is an arrangement in which individuals consider the amount, level, value, worth, quality, or success of the products or outcomes of learning of peers of similar status".

The PA section focused on seven topics (details in Appendix).

(a) PA usage. The three first questions established whether teachers had used PA at all, how frequently they used it, if they had, and for how many years they had implemented it. These allowed identification of valid cases and the distinction between usual and occasional users and between recent and established users.
(b) Valence of the experience. This question established whether participants considered their experience of PA to be positive or negative.
(c) PA privacy format. This question established whether the assessor and assessees were anonymous or not and how feedback was given, if present. These responses established the degree of privacy in PA.
(d) PA accuracy. Six questions in total were used to establish whether teachers believed students were or were not accurate and why (open question) and follow-up as to whether teachers considered if students tended to over-score, under-score, or accurately score.
(e) PA official weight. Three items explored whether teachers would (a) allow the PA score to be included in the course grade, (b) their reasons, and (c) what percentage of the grade would be dependent on PA.
(f) PA disadvantages. To decide which problems with PA to include in the survey a list with five problems was created. This list was analyzed by three formative assessment scholars. One of them proposed an additional problem, but all three agreed the list was compelling. Therefore, these six most commonly reported problems with PA were listed and participants were instructed to choose as many as they considered relevant. Additionally, an open-ended "other" category allowed teachers to report problems not anticipated by the list.
(g) PA advantages. The same procedure as mentioned was followed here: five commonly reported advantages for PA were listed, analyzed by the three experts, included in the survey and participants were instructed to choose as many as they considered relevant. An open-ended "other" category allowed teachers to report additional advantages.

A pilot of the survey was conducted before administration. An expert in formative assessment filled out the questionnaire in front of the first author following a think-aloud protocol. That input was used to revise some of the items. The revised survey was then evaluated with think-aloud procedures by a primary and a secondary teacher. Finally, seven teachers filled out the revised survey on their computers to simulate online administration. These teachers sent their feedback about the survey itself (e.g., comprehension problems, length, etc.). Additionally, to validate responses in the actual administration, the last question was an open question in which participants were invited to express any concerns or suggestions they had regarding the survey itself or the whole study.

Procedure

Requests for participation were sent to the person in charge of each education centre (e.g., principal in primary school) who was asked to alert the centre teachers to the opportunity to participate in an online survey. The request for participation contained details about the study (i.e., purpose, confidentiality assurances, and URL for the Google survey).

Analysis

Basic descriptive statistics (i.e., mean, standard deviation, and frequency) by level of employment were utilized to answer RQ1. Additionally, content analysis were used for the two "Why" open-ended questions asking about PA official weight and accuracy. The first author plus a research assistant read 30 % of the answers in a random order. Then they created categories to group the answers, which were then used to independently categorize all answers. A random sample of 30 % of the items in each category was used to calculate inter-rater agreement Cohen's kappa. In the scarce occasions where there was a discrepancy, the first author coding was taken due to a higher expertise in PA. For the open-ended accuracy item the agreement was high ($\kappa = .93$) as was the agreement ($\kappa = .95$) for the official weight of PA item.

Since there were four items exploring the pros and cons of PA each, a multiple indicator, factor analytic approach was used to simplify the dimensionality of the items (RQ1). A two-factor solution (i.e., advantages and disadvantages) was tested for quality of fit to the data and admissibility with confirmatory factor analysis (CFA).

Since the goal of RQ2 was to understand how teacher beliefs influence self-reported PA implementation, a causal-correlational approach was used to identify the relationship of latent factors and manifest item variables to each other and their contribution, if any, to PA usage. Structural equation modeling (SEM), based on the exploratory analyses conducted for RQ1 was used to propose a structure of relations among factors to predict self-reported frequency of PA use. SEM is preferable because it generates multiple indices that indicate how closely the model fits the data and because it incorporates latent factors with manifest variables into a causal path. This provides a more sophisticated evaluation of the relationship of factors to each other (Bollen 1989; Borsboom 2006).

To answer RQ3 about differences between educational levels, a multigroup invariance test of the structural model was utilized. This allows the identification of whether level of teaching produced a statistically equivalent set of parameter estimates in the overall model reported in RQ2. Lack of statistical equivalence of regression weights or lack of identical configuration of paths in the model itself indicates responses for each group need to be treated separately. Support for equivalence is found when the difference in the comparative fit index (ΔCFI) is not more than .01 (Vandenberg and Lance 2000).

Although the questionnaire used binary and ordinal variables, the ML estimator in Amos (IBM 2011 was utilized for all EFA, CFA, and SEM procedures. Model fit in CFA and SEM was determined by inspection of multiple indices (Hu and Bentler 1999; Fan and Sivo 2007); conventional standards for good fit are statistically non-significant probability ($p > .05$) for the ratio of χ^2/df (Marsh, Hau, and Wen 2004), CFI and gamma hat (γ) >.95 (Fan and Sivo 2007); root mean square error of approximation (RMSEA) <0.05 (Hu and Bentler 1999), and the standardized root mean residual (SRMR) <.06 (Hu and Bentler 1999).

Results

RQ1—What do teachers think about the various aspects of PA?

As there is a significant amount of data collected to answer RQ1 the results will be organized in four sub-research questions.

RQ1a—Do Spanish teachers report using PA, with what frequency, what is their experience, and what is the preferred PA privacy format?

Overall, teachers believed in students' participation in assessment (Table 2), a crucial prerequisite to implement PA in the classroom. University teachers were the most reluctant with only a 55 % supporting the idea, while primary and secondary teachers largely supported it. Self-reported use of PA in the classroom was highest in primary (68 %) and secondary (55 %) and lowest among university teachers (37 %). However, frequency of PA use was generally low, with "occasionally" being the most chosen option for primary and secondary teachers, and second most frequent among university teachers. Most teachers who had used PA reported having positive or neutral experiences.

Overwhelmingly, teachers reported that both the assessor and assessee would not be anonymous in PA (Table 3). A large majority of teachers indicated that PA feedback would not be given individually, nor in groups, and not in the classroom.

Table 2 Teachers' uses of peer assessment

Question and response category	Primary ($N=441$)		Secondary ($N=689$)		University ($N=155$)	
	n	%	*n*	%	*n*	%
Necessity that students participate in assessment						
Not necessary	47	10.7	104	15.1	45	29
Sometimes/depends	30	6.8	62	9	16	10.4
Yes/yes I already do that	364	82.5	516	74.9	85	54.8
Other	0	0	7	1	9	5.8
Use in my courses						
Yes	300	68.2	378	54.9	58	37.4
No	140	31.8	310	45.1	97	62.6
Frequency of use						
Daily	11	2.5	3	.4	1	.6
Weekly	32	7.3	15	2.2	4	2.6
Monthly	42	9.6	46	6.7	8	5.2
Occasionally	220	50.1	329	47.8	47	30.3
Never	134	30.5	295	42.9	95	61.2
Experience with PA	$n=300$		$n=400$		$n=56$	
Negative	15	5	22	5.5	4	7.1
Neutral	69	23	123	30.8	12	21
Positive	216	72	255	63.7	40	71.4

Table 3 Peer assessment format

Format and response	Primary ($N = 441$)		Secondary ($N = 690$)		University ($N = 155$)	
	n	%	*n*	%	*n*	%
The assessee is anonymous						
No	407	92.3	626	90.7	146	94.2
Yes	34	7.7	64	9.3	9	5.8
The assessor is anonymous						
No	416	94.3	638	92.5	142	91.6
Yes	25	5.7	52	7.5	13	8.4
Feedback is given individually						
No	369	83.7	593	85.9	136	87.7
Yes	72	16.3	97	14.1	19	12.3
Feedback given in working groups						
No	287	65.1	494	71.6	120	77.4
Yes	154	34.9	196	28.4	35	22.6
Feedback is given in classroom						
No	340	77.1	581	84.2	136	87.7
Yes	101	22.9	109	15.8	19	12.3

This leaves it unclear as to how teachers would implement feedback to students from PA. Since feedback in working groups had the highest rating, it is possible that teachers expect students to give feedback from PA interactively with their individual peer, without input from the teacher, although this is a highly speculative explanation.

RQ1b—Do they consider PA accurate? Why?

Table 4 provides details of teachers' evaluation of student accuracy in PA. Generally, two-fifths of primary and secondary teachers agreed that students were accurate, with just a third of university teachers agreeing. About one third of primary teachers, two fifths of secondary teachers, and nearly half of the university teachers believed students were not accurate in PA. As teaching level increased the proportion of teachers thinking students under-scored declined, while the proportion of teachers believing students tended to over-score increased. These two options accounted for a very large proportion of all teachers (69 % primary teachers, 75 % secondary teachers, and 72 % of university teachers). In sum, PA accuracy is a real concern for teachers, especially at more advanced educational levels.

From the open-ended question, it was possible to identify eight major reasons teachers gave for why they thought students were or were not accurate in PA (Table 5). Nearly three quarters of all responses from primary and secondary teachers and almost two thirds of university teacher responses were reasons for inaccuracy in student PA. The three main reasons teachers did *not* consider PA to be accurate were: (a) the effect of interpersonal relationships (e.g., friendship scoring), (b) students being unrealistically demanding, and (c) students lacking expertise. Accuracy was identified as possible in only about 10 % of answers only when students received

Table 4 Teachers' perceptions about peer assessment accuracy

	Primary (*N* = 440)		Secondary (*N* = 688)		University (*N* = 154)	
	n	%	*n*	%	*n*	%
Are students accurate when peer assessing						
No	162	36.8	288	41.9	71	46.1
Yes	192	43.6	269	39.1	50	32.5
Depends (unspecified)/sometimes	53	12	74	10.8	14	9.1
NK/NA	8	1.8	30	4.4	19	12.3
Yes, once they learn	4	.9	7	1	0	0
They are too hard underscoring	13	3	9	1.3	0	0
Depends on relationship with assessee	8	1.8	11	1.6	0	0
Students' tendency in PA scoring	*n* = 248		*n* = 356		*n* = 85	
Underscore	128	51.6	143	40.2	19	22.4
Overscore	44	17.7	123	34.6	42	49.4
Both	14	5.6	24	6.7	3	3.5
They are accurate	8	3.2	11	3.1	4	4.7
Depends (unspecified)	11	4.4	14	3.9	5	5.9
Depends on relationship with the assessee	36	14.5	30	8.4	6	7.1
They are accurate if they have assessment criteria	2	.8	2	.08	5	5.9
They are hard underscoring	0	0	1	.003	0	0
Other	5	2	8	2.2	1	1.2

adequate training and/or criteria for PA. Hence, significant concerns existed throughout the teaching profession about the accuracy of peer assessment.

Table 5 Reasons for teachers' beliefs on PA accuracy

Categories	Primary (*N* = 176)		Secondary (*N* = 295)		University (*N* = 54)	
	n	%	*n*	%	*n*	%
Not accurate reasons						
Interpersonal relationships	48	27.27	87	29.49	13	24.07
Students are too demanding	47	26.70	61	20.68	9	16.67
Students are not experts	9	5.11	16	5.42	8	14.81
Students are not objective	7	3.98	24	8.14	4	7.41
Students are competitive with their peers	6	3.41	16	5.42	0	0
Students lack of maturity	9	5.11	7	2.37	0	0
Sub-total	126	71.58	211	71.52	34	62.96
Accurate reasons						
If students receive training and/or criteria	21	11.93	24	8.14	4	7.41
Students are objective	11	6.25	19	6.44	3	5.56
Sub-total	32	18.18	43	14.58	7	12.97
Other	18	10.23	41	13.90	13	24.07

RQ1c—Would teachers let a percentage of their course grade depend on PA score? Which percentage? Why?

Teachers were almost equally split between using or not using PA scores towards their course final grade (Table 6). Few teachers provided an estimate of how much weight towards the final course grade they would give to PA (i.e., 26 % primary, 37 % secondary, and 21 % university). Both primary and secondary teachers indicated that it would count on average for 18 % of the final score, while university teachers would allow it to count for only 5 %.

Generally, more of the reasons given related to *not* including PA in course grades, except for higher education teachers (Table 7). Bias to do with interpersonal relationships was the most common objection followed by a similar concern that students would not be objective, though higher education teachers were more concerned about the general unfairness of PA. In contrast, positive reasons for including PA rose with increase in teaching level. Higher education teachers saw the responsibility that PA engenders as warranting its use, while secondary teachers emphasized the increased responsibility for learning inherent in PA. This latter reason accounted for half of the positive reasons given by primary teachers. These self-provided reasons more or less align with the forced choice responses and may shed light on the results of the structural equation model.

RQ1d—What are the main advantages and disadvantages of PA?

In Table 8 the reported advantages and disadvantages are presented. In regard to advantages, making students more responsible for their learning was the dominant reason for using PA, with three reasons being endorsed by 40–50 % of participants (i.e., help with group work, helping learning, and help in detecting problems). Less than 10 % considered PA would save teachers time. On the other hand, around half of teachers saw low reliability and student mistrust of PA scoring as disadvantages to PA, followed by a third to two fifths who considered PA caused problems in classroom climate. Otherwise, the three remaining disadvantages were selected by less than 10 % of participants.

Confirmatory factor analysis of the positive reasons for using PA and the negative reasons for avoiding PA were run in a 2-factor inter-correlated model. It was found that the three negative items with very low frequency of selection did not have statistically significant paths from the negative factor. After their removal a two-factor model (Fig. 1) with acceptable to

Table 6 Official weight PA

	Primary ($N=440$)		Secondary ($N=688$)		University ($N=154$)	
	n	%	*n*	%	*n*	%
Would you let a percentage of your course grade depends on PA score?						
No	226	51.6	338	49.3	63	40.6
Maybe/depends	20	4.6	52	7.6	15	9.6
Yes/yes I already do that	192	43.8	295	43.1	77	49.7
Percentage PA course score	$n=77$		$n=146$		$n=12$	
What percentage of your course would you allow to depend on PA?	M = 18.39; SD = 13.42; range = 0–50		M = 18.12; SD = 17.04; range = 0–100		M = 5.00; SD = 5.64; range = 5–20	

Table 7 Reasons to include or not PA score as part of the final course grade

Categories	Primary (*N* = 64)		Secondary (*N* = 106)		University (*N* = 14)	
	n	%	*n*	%	*n*	%
Non-inclusion reasons						
PA is biased by students relationship	18	28.13	25	23.58	1	7.69
Students lack of objectivity	11	17.19	14	13.21	0	0
PA is not reliable	1	1.56	4	3.77	1	7.69
Respect among peers for their comments	3	4.69	1	0.94	1	7.69
PA requires hard work and preparation	2	3.13	5	4.72	1	7.69
Assessment is the teachers' duty and expertise	4	6.25	9	8.49	0	0
Generally unfair	2	3.13	3	2.83	2	15.38
Sub-total	41	64.06	61	57.55	6	42.86
Inclusion reasons						
PA is part of the learning process (e.g., assuming more responsibility)	10	15.63	29	27.36	0	0
Objectivity (whatever it means)	3	4.69	1	0.94	0	0
PA is reliable	2	3.13	1	0.94	2	15.38
Effort (whatever it means)	4	6.25	3	2.83	4	30.77
Only for less important tasks	3	4.69	8	7.55	0	0
Provides new points of view	1	1.56	3	2.83	1	7.69
Sub-total	23	35.94	45	42.45	7	50
Other					1	7.69

Table 8 Teachers' reported endorsement of PA advantages and disadvantages by level of employment

Category	Primary (*N* = 441)		Secondary (*N* = 690)		University (*N* = 155)	
	n	%	*n*	%	*n*	%
Advantages						
Students more responsible	343	77.8	516	74.8	113	72.9
Help in group work	206	46.7	305	44.2	62	40
Students learn	202	45.8	277	40.1	79	51
Detect problems	187	42.4	286	41.4	59	38.1
Save time for teachers	30	6.8	34	4.9	9	5.8
Other	1	.22	2	.29	0	0
Disadvantages						
Low reliability	232	52.6	384	55.7	85	54.8
Students do not trust PA scores	234	53.1	335	48.6	71	45.8
Problems classroom climate	149	33.8	235	34.1	66	42.6
Loss time more than saves	31	7	89	12.9	14	9
Students do not learn via PA	12	2.7	46	6.7	9	5.8
Loss teacher's authority	7	1.6	15	2.2	2	1.3
Other	3	.68	4	.58	1	.64

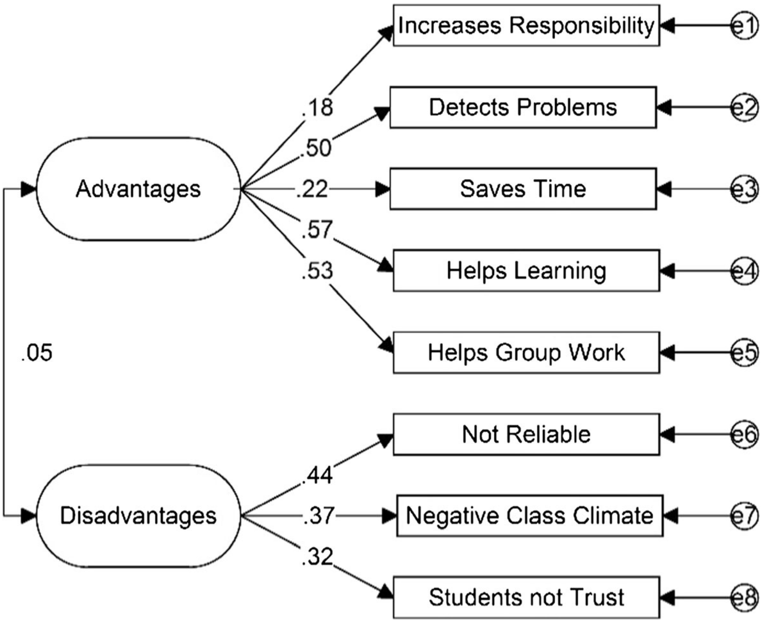

Fig. 1 CFA model of advantages and disadvantages in the use of PA

good fit was found (χ^2 = 39.16, *df* = 19, χ^2/*df* = 2.06 (*p* = .15); CFI = .92; gamma hat = .99; RMSEA = .038 (90 % confidence interval (CI) = .021–.054); SRMR = .035). The inter-correlation was close to zero (*r* = .05) showing that these two constructs were independent. Mean scores for the positive factor (M = .48, SD = .25) were not statistically significant by teaching level ($F_{(2,\ 748)}$ = 1.02, *p* = .36) and for the negative factor (M = .44, SD = .32) were trivially different ($F_{(2,\ 748)}$ = 3.34, *p* = .04, eta^2 = .009). This indicates that the teacher groups did not differ in their average frequency of selecting these reasons for using or avoiding PA.

RQ2—How do these beliefs influence self-reported PA implementation?

A structural equation model (acceptable to good fit: χ^2 = 47.12, *df* = 19, χ^2/*df* = 2.48 (*p* = .12); CFI = .95; RMSEA = .044 (90 % CI = .029–.061); SRMR = .036) identified that three advantages of PA (i.e., detect problems, help with group work, and student learning) were predicted by teachers' PA experience (Fig. 2). This general factor consisted of previous positive experience, previous use of PA, willingness to include PA as part of final grade, and belief in student participation in assessment. Together, these perceptions and self-reported uses of PA predicted greater frequency of using PA, with a large effect (R^2 = .28).

Together, the structural equation model indicates that PA classroom usage is supported by beliefs in its positive contribution to student learning combined with positive previous experiences. None of the negative factor obstacles could be fit to the model, suggesting that teachers' awareness of the difficulties, does not determine their choice to use PA in the classroom setting.

RQ3—What differences exist between educational levels?

In the descriptive analysis reported earlier, higher education teachers' PA implementation was lower (Table 2), they were less positive about the accuracy of PA (Table 4), and they would

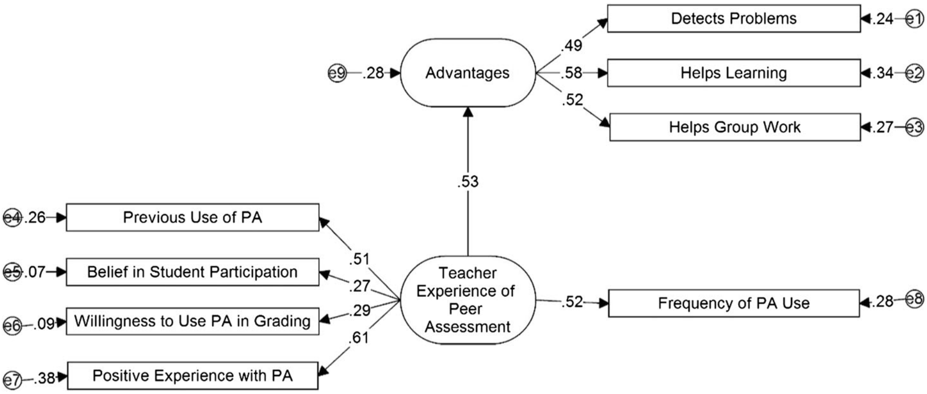

Fig. 2 SEM model of factors influencing frequency of PA implementation

only let a small percentage of the course grade depend on PA score (see Table 6) compared with the K-12 teachers. Therefore, evaluating the structural equation model by teachers' level of employment was conducted. A multi-group confirmatory factor analysis was attempted with all three groups, but unsurprisingly, the university group with its small sample size ($n = 55$) was not configurally equivalent to the main model. For example, among the university teachers, having anonymous peer assessors had a substantial contribution to frequency of use ($\beta = .41$, $p < .001$), but this was not significant for the K-12 teachers.

The two-group analysis of primary and secondary teachers required fixing the error variance of two items to a small positive value (Chen et al. 2001). The model did not have equivalent regression weights (ΔCFI >.01). Hence, the Fig. 2 model derived from the responses of all participants did not have equivalent regression weights for the three groups of teachers.

Inspection of the standardised regression weights of the model (λ) for the three groups shows that four paths differed by considerable margins (i.e., $\Delta\lambda$ >.10) across the three groups (Table 9). Of these paths, the most striking difference was seen in willingness to use PA in

Table 9 Model path weights and variance explained by group

Factor	Combined		Primary		Secondary		University	
	λ	R^2	λ	R^2	λ	R^2	λ	R^2
PA advantages								
Detects problems[a]	0.49	.24	0.99	.98	0.99	.98	0.99	.98
Helps learning	0.58	.34	0.25	.06	0.29	.09	0.24	.06
Helps group work	0.52	.27	0.27	.07	0.24	.06	0.31	.10
Use of PA								
PA advantages	0.52	.28	0.10	.01	0.13	.02	0.12	.01
Previous use of PA[a]	0.51	.26	0.94	.88	0.97	.94	0.85	.72
Belief in student participation	0.27	.07	0.07	.01	0.15	.02	0.09	.01
Willingness to use PA in grading	0.29	.09	0.17	.03	0.17	.03	−0.75	.57
Positive experience with PA	0.61	.38	0.35	.12	0.32	.10	0.03	.00
Frequency of PA use	0.52	.28	0.36	.13	0.44	.20	0.24	.06

[a] Residual variance value constrained to .005 in multi-group analysis

grading, which had weakly positive values for primary and secondary teachers and very strong negative value for university teachers, clearly indicating quite strikingly different evaluation of the role PA can play in grading by level of teaching. This especially confirms that reporting of the results had to be done separately for each group.

The amount of variance explained in the frequency of using PA (R^2) varied considerably by teaching level, with the greatest amount seen among the secondary teachers and least at the university level. These further suggest strong differences in how PA is valued and used at each level.

Discussion

The aim of this paper was to explore teacher perceptions of PA practices at all educational levels in order to better understand the effect of their beliefs on self-reported uses of this assessment practice. This study has used a large sample of K-12 teachers and a smaller one from higher education in a previously little studied non-English speaking jurisdiction (Spain).

Use of PA

Consistent with previous studies (Lynch and Golen 1992; Noonan and Duncan 2005), the use of peer assessment was at best an occasional practice, even though PA was viewed positively. Given how little training teachers reported having in assessment, let alone PA, this occasional use may be entirely appropriate. In contrast to those studies, this sample of Spanish teachers was much larger, generating a more robust picture of teacher thinking about the use of PA.

Overall, the teachers were positive about the ability of PA to help students take responsibility for their own learning and, as expected, their beliefs about PA had a statistically significant and robust effect on their self-reported use of PA. Consistent with our expectations, the structural model shows greater use of PA is associated with focusing on the positive learning advantages of PA. Contrary to our expectations, the negative reasons or obstacles teachers were able to identify for not using PA did not meaningfully explain their willingness to use PA. Furthermore, the notion of PA saving time for teachers was not endorsed, nor did it predict greater use of PA. This belief is reassuring since it is consistent with research that shows effective PA requires more effort from teachers (Panadero 2016). Hence, it would seem relatively uncontroversial to conclude that teachers like the idea of PA, struggle somewhat with inherent difficulties, and that their self-reported use depends largely on previous positive experiences.

Anonymity

This group of teachers believed predominantly in using anonymous versions of PA. This could be a positive result because anonymity has been found to reduce the impact of PA on interpersonal conflict and tension among students (Vanderhoven et al. 2015). However, contrary to our expectations, anonymous modality did not have a statistically significant relationship to PA frequency in the structural model, except for anonymous assessors among university teachers. Anonymous assessors at the university level seems consistent with the relatively low official weight endorsed for PA at the university level and may function to ensure honesty and accuracy in peer appraisals.

Nevertheless, PA anonymity is not yet conclusively positive for interpersonal variables, because knowing the feedback giver or being known as a peer assessor may help students learn more. When students know and potentially trust the sources of PA (i.e., non-anonymity condition), more face to face interaction can occur around the learning topic (Panadero 2016). Teachers seemed to be aware that students could be resistant to PA. A possible explanation is that if PA has been implemented purely in a summative grading approach with little feedback, then negative interpersonal relationships among students could easily arise (Panadero 2016). The teachers might have concluded from their experience that students dislike PA, instead of considering it was the risks in the form of PA being used that were creating the problems. Therefore, greater attention among teachers to minimizing the use of PA for grading and increasing the learning benefits is needed.

Depth of PA

The teachers in this study did not clarify whether the PA they were referring to involved significant amounts of peer feedback. Thus, we do not know how deep their implementation was. It is possible that the teachers implemented rather superficial approaches to PA, involving mostly peer scoring or unguided peer comments instead of PA with scaffolded and extensive peer feedback. If that is the case, then these results may not be as encouraging as they currently seem. Being positive about and supportive of less effective forms of PA may only be a first step towards effective teacher PA practice. But it may also make teachers immune to further development in their use of PA.

Accuracy in PA

The teachers were clearly concerned about the accuracy of PA with a high proportion concerned that students tended to over-score. However, at the same time these concerns over reasons not to use PA, did not play a meaningful role in predicting self-reported use of PA. Indeed, despite many teachers indicating that PA had negative aspects, the positive advantages of using PA for greater learning determined whether or not PA was being used. This result suggests that the training of teachers needs to counter negative perceptions based on empirical research that shows students can be reliable sources under appropriate conditions (Falchikov and Goldfinch 2000; Topping 2003). But more importantly, teachers need to become persuaded of the greater importance of the benefits of PA rather than focus on its problems. Therefore, it is clear that training teachers to use PA also needs to focus on how teachers can best implement PA to ensure students can be accurate peer assessors.

Official weight of PA

It is clear that very similar proportions of all teachers across levels would allow PA to count towards grades but the allowed weight was much greater for K-12 teachers than among university teachers. Contrary to our expectations, giving some official weight was a small but positive predictor of increased use of PA. It may be that giving weight is seen as motivating students to greater attention and effort since their peer evaluation counts. However, this could be problematic because, without proper supervision to minimize friendship collusion or hatred marking, there are opportunities for construct irrelevance in such marking. From a purely learning point of view, any official weight may be counter-productive, whereas, as long as students are instrumentally and strategically motivated, not giving it official weight may reduce effort and attention. Thus, a dilemma exists as to how to achieve the positive effects of PA and how to minimize student resistance or corruption with PA. This could be achieved by helping teachers design more intensive PA practices, such as allowing students to give extensive peer feedback so as to decrease negative interpersonal effects (Panadero 2016).

Educational level differences

Consistent with our expectations and previous research on self-assessment use (Panadero et al. 2014), use of PA was higher in the K-12 sector than in higher education, although among those using it the perceptions were relatively equally positive. Indeed, the structural model was different, contrary to expectations, for higher education teachers, though this may be due to the relatively small sample size. Whether this difference is a problem cannot be easily resolved. Higher education should be an optimal educational level in which to implement PA, given the relative maturity and competence of such young adults, and since considerable reliability in PA has been demonstrated in higher education. Furthermore, higher education students upon graduation and entry into work are supposed to, not only be able to evaluate their own work, but also that of their future work colleagues and, concomitantly, have their work evaluated by peers.

It may be that this negative trend in higher education is not a function of teacher beliefs but rather a natural consequence of the increasingly constrained nature of teaching as students progress through schooling, suggesting that there might be robust systemic reasons for not using PA in higher education; though the following interpretations have to be treated as speculative. The relative low use of PA in higher education may be a rational response to conditions that do not exist in K-12 schooling, as suggested for self-assessment by Panadero et al. (2014). For example, European universities under the influence of processes such as the Bologna Declaration seem to constrain assessment practices so that they are transparent and consistent and also ensure that the performance of an individual learner can be adequately assured. Hence, policies often prevent a large proportion of course grades being based on group work and tend to give greater weight to performance under formal examination conditions to ensure that grades reflect the individual's capability and not that of a group member or some external source as is possible in take-home coursework.

Further, courses in university are generally much briefer and with many fewer contact hours than K-12 courses, meaning that opportunities to conduct group work are much less. Additionally, university classes are far larger than primary and secondary classes, making formative assessment a more complex enterprise (e.g., ensuring high trust between peer assessor and assessee). In contrast, K-12 teachers seem to have greater flexibility in school-based assessment to include a wider variety of assessment practices, including PA which can be managed more easily since there are many more opportunities for peer interaction and interpersonal knowledge in courses that meet 4 h/week for at least 30 weeks/year.

It is also possible that university teachers do hold learning and assessment beliefs closer to a transmission of knowledge belief rather than a development of learning competences view, which would make student involvement in assessment less likely (Tan 2012). It may also be that teachers in higher education have less pedagogical and assessment training than their primary and secondary counterparts and are, thus, under-equipped to use PA. Nonetheless, the results clearly indicate that the variables included in the study were insufficient to explain the usage of PA among Spanish higher education teachers; this suggests other factors are necessary to properly understand PA in higher education.

Implications

The implications for practice of this study are significant. Because a positive relationship between PA experience and awareness of its advantages are meaningful predictors of self-reported use, training programs that build teacher competence in PA are needed. A crucial lever may be getting teachers to practice PA themselves with other teachers in a professional development program so as to gain expertise prior to their giving students training in PA. Such practice would give teachers greater awareness of the interpersonal dynamics and challenges in giving and receiving feedback and evaluation from a peer. Such professional development needs to take into account the guidelines already available (e.g., Topping 2003). Nonetheless, it is also clear that policy frameworks and systemic realities shape the possibilities of implementing PA. It is not sensible to require more PA that counts when institutional policies prevent this. Thus, greater understanding of PA as a powerful pedagogical practice to ensure students develop competencies needed in life beyond school should be a key goal in developing teacher beliefs about PA.

Limitations

The major limitation of this study is its self-reported nature. Therefore, this study does not use a test of knowledge or behavior; it is a measure of teacher perceptions or beliefs. Hence, responses may reflect some elements of social desirability. Perhaps quite different results would arise were the students of these teachers surveyed or their classrooms observed. However, the consistency between responses to fixed-format and open-ended questions is somewhat reassuring that at least internal consistency is evident. Nonetheless, future studies would do well to triangulate teacher espoused beliefs with their enacted behaviors.

Another matter of concern is the use of the vague quantity "occasionally" in the response scale for how often teachers use PA. Clearly, there are memory problems within individuals in recalling how often a practice has been implemented and there is variability between individuals in how such a vague frequency is understood (Schacter 1999). Future research needs to find a more robust mechanism for establishing commonality across individuals and mitigating faulty memory problems. Solutions could include use of an agreement rating scale or restricting the memory to a fixed time period (e.g., in the last teaching semester).

Conclusions

While in general, teachers valued the potential of PA, quite different relationships among the constructs were seen according to teaching level. This suggests quite different responses, which are required to change either policy or practice constraints or teacher beliefs and values at each level of schooling. Without infrastructure support through policy, professional development, and resources, it is highly unlikely the positive view of PA seen in these results will be converted into actual usage, meaning the potential life and academic gains of PA will not become activated.

Acknowledgments Research funded by personal grant to Ernesto Panadero under Ramón y Cajal framewok (RYC-2013-13469).

Appendix

Table 10 Teachers' beliefs about students peer assessment survey

Definition: "peer assessment is an arrangement in which individuals consider the amount, level, value, worth, quality, or success of the products or outcomes of learning of peers of similar status" (Topping 1998).

No.	Question	Response format
1	Have you used peer assessment in your courses?	2-point agreement
2	With what frequency?	5-point rating + other
3	How long have you been using it?	Frequency (years)
4	In case you have use peer assessment, how would you consider your experience?	3-point rating
5	In case you use peer assessment, what modalities of peer assessment have you used? (a) The assessee was anonymous. (b) The assessor was anonymous. (c) Feedback was given in an individual basis. (d) Feedback was given in the working groups. (e) Feedback was given to all the classroom	Y/N in each of the five categories.
6i	Do you consider students accurate when assessing a classmate?	3-point rating
6ii	Why?	Open question
7	If you consider that they are not accurate, do you think they tend to…?	2-point rating + other
8i	Would you let a percentage of your course grade would depend on the peer assessment score?	3-point rating + Other
8ii	Why?	Open question
8iii	In case you would let, what exact percentage would you allow?	Percentage, 0–100 %
9i	Problems: reliability	Y/N
9ii	Problems: creates problems for the teacher's authority	Y/N
9iii	Problems: it causes more loss of time that the one it saves	Y/N
9iv	Problems: creates problems in the classroom group as they have to assess one another	Y/N
9v	Problems: students don't trust the score given by their classmates	Y/N
9vi	Problems: it does not enhance students' learning	Y/N
9vii	Problems: other	Open question
10i	Advantages: students are more conscious and responsible for their learning by reflecting on their classmates work	Y/N
10ii	Advantages: detection and correction of problems	Y/N
10iii	Advantages: saves time for the teacher	Y/N
10iv	Advantages: students learn using that strategy	Y/N
10v	Advantages: helps on group work	Y/N
10vi	Advantages: other	Open question

References

Ajzen, I. (2005). *Attitudes, personality and behavior* (2nd ed.). New York: Open University Press.

Barnes, N., Fives, H., & Dacey, C. M. (2015). Teachers' beliefs about assessment. In H. Fives & M. Gregoire Gill (Eds.), *International Handbook of Research on Teacher Beliefs* (pp. 284–300). New York: Routledge.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice, 5*(1), 7–74.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: John Wiley & Sons.

Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika, 71*(3), 425–440. doi:10.1007/s11336-006-1447-6.

Bryant, D. A., & Carless, D. R. (2010). Peer assessment in a test-dominated setting: Empowering, boring or facilitating examination preparation? *Educational Research for Policy and Practice, 9*(1), 3–15. doi:10.1007/s10671-009-9077-2.

Carvalho, A. (2012). Students' perceptions of fairness in peer assessment: Evidence from a problem-based learning course. *Teaching in Higher Education, 18*(5), 491–505. doi:10.1080/13562517.2012.753051.

Chen, F., Bollen, K. A., Paxton, P., Curran, P. J., & Kirby, J. B. (2001). Improper solutions in structural equation models: Causes, consequences, and strategies. *Sociological Methods & Research, 29*(4), 468–508.

Dochy, F., Segers, M., & Sluijsmans, D. (1999). The use of self-, peer- and co-assessment in higher education. *A review. Studies in Higher Education, 24*(3), 331–350. doi:10.1080/03075079912331379935.

Falchikov, N. (1995). Peer feedback marking: Developing peer assessment. *Innovations in Education & Training International, 32*(2), 175–187. doi:10.1080/1355800950320212.

Falchikov, N., & Goldfinch, J. (2000). Student peer assessment in higher education: A meta-analysis comparing peer and teacher marks. *Review of Educational Research, 70*(3), 287–322.

Fan, X., & Sivo, S. A. (2007). Sensitivity of fit indices to model misspecification and model types. *Multivariate Behavioral Research, 42*(3), 509–529.

Fives, H., & Buehl, M. M. (2012). Spring cleaning for the "messy" construct of teachers' beliefs: What are they? Which have been examined? What can they tell us? In K. R. Harris, S. Graham, & T. Urdan (Eds.), *APA Educational Psychology Handbook: Individual Differences and Cultural and Contextual Factors* (Vol. 2, pp. 471–499). Washington, DC: APA

Gao, M. (2009). Students' voices in school-based assessment of Hong Kong: A case study. In D. M. McInerney, G. T. L. Brown, & G. A. D. Liem (Eds.), *Student perspectives on assessment: What students can tell us about assessment for learning* (pp. 107–130). Charlotte, NC: Information Age Publishing.

Harris, L. R., & Brown, G. T. L. (2013). Opportunities and obstacles to consider when using peer- and self-assessment to improve student learning: Case studies into teachers' implementation. *Teaching and Teacher Education, 36*(0)), 101–111. doi:10.1016/j.tate.2013.07.008.

Hu, L.-T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*(1), 1–55.

Hwang, G. J., Hung, C. M., & Chen, N. S. (2014). Improving learning achievements, motivations and problem-solving skills through a peer assessment-based game development approach. *Educational Technology Research and Development, 62*(2), 129–145. doi:10.1007/s11423-013-9320-7.

IBM. (2011). Amos [computer program] (version 20, Build 817). Meadville, PA: Amos Development Corporation

Ion, G., & Cano, E. (2011). Assessment practices at Spanish universities: From a learning to a competencies approach. *Evaluation & Research in Education, 24*(3), 167–181. doi:10.1080/09500790.2011.610503.

Kim, M., & Ryu, J. (2013). The development and implementation of a Web-based formative peer assessment system for enhancing students' metacognitive awareness and performance in ill-structured tasks. *Educational Technology Research and Development, 61*(4), 549–561. doi:10.1007/s11423-012-9266-1.

Lingard, B., & Lewis, S. (2016). Globalization of the Anglo-American approach to top-down, test-based educational accountability. In G. T. L. Brown & L. R. Harris (Eds.). *Handbook of Social and Human Conditions in Assessment* (pp. 1–30). New York: Routledge.

Lynch, D. H., & Golen, S. (1992). Peer evaluation of writing in business communication classes. *Journal of Education for Business, 68*(1), 44–48. doi:10.1080/08832323.1992.10117585.

Marsh, H. W., Hau, K.-T., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. *Structural Equation Modeling, 11*(3), 320–341.

Nicol, D. J., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in Higher Education, 31*(2), 199–218.

Noonan, B., & Duncan, C. R. (2005). Peer and self-assessment in high schools. *Practical Assessment, Research & Evaluation, 10*(17). http://pareonline.net/getvn.asp?v=10&n=17.

Panadero, E., Brown, G. T. L., & Courtney, M. G. R. (2014). Teachers' reasons for using self-assessment: A survey self-report of Spanish teachers. *Assessment in Education: Principles, Policy & Practice, 21*(3), 365–383. doi:10.1080/0969594X.2014.919247.

Panadero, E., Romero, M., & Strijbos, J. W. (2013). The impact of a rubric and friendship on construct validity of peer assessment, perceived fairness and comfort, and performance. *Studies In Educational Evaluation, 39*(4), 195–203. doi:10.1016/j.stueduc.2013.10.005.

Panadero, E. (2016). Is it safe? Social, interpersonal, and human effects of peer assessment: a review and future directions. In G. T. L. Brown & L. R. Harris (Eds.), *Human factors and social conditions of assessment.* New York: Routledge.

Panadero, E., Jonsson, A., & Strijbos, J. W. (2016). Scaffolding self-regulated learning through self-assessment and peer assessment: Guidelines for classroom implementation. In D. Laveault & L. Allal (Eds.), Assessment for learning: meeting the challenge of implementation. Springer, In press.

Peterson, E. R., & Irving, S. E. (2008). Secondary school students' conceptions of assessment and feedback. *Learning and Instruction, 18*(3), 238–250.

Reinholz, D. L. (2015). The assessment cycle: A model for learning through peer assessment. *Assessment & Evaluation in Higher Education*, 1–15. doi: 10.1080/02602938.2015.1008982

Remesal, A. (2007). Educational reform and primary and secondary teachers' conceptions of assessment: The Spanish instance, building upon Black and Wiliam (2005). *Curriculum Journal, 18*(1), 27–38. doi:10.1080/09585170701292133.

Remesal, A. (2011). Primary and secondary teachers' conceptions of assessment: A qualitative study. *Teaching and Teacher Education, 27*(2), 472–482. doi:10.1016/j.tate.2010.09.017.

Schacter, D. L. (1999). The seven sins of memory: Insights from psychology and cognitive neuroscience. *American Psychologist, 54*(3), 182–203.

Spandorfer, J., Puklus, T., Rose, V., Vahedi, M., Collins, L., Giordano, C., & Braster, C. (2014). Peer assessment among first year medical students in anatomy. *Anatomical Sciences Education, 7*(2), 144–152. doi:10.1002/ase.1394.

Tan, K. H. K. (2012). *Student self-assessment. Assessment, learning and empowerment.* Singapore: Research Publishing.

Topping, K. J. (1998). Peer assessment between students in colleges and universities. *Review of Educational Research, 68*(3), 249–276.

Topping, K. J. (2003). Self and peer assessment in school and university: Reliability, validity and utility. In M. Segers, F. Dochy & E. Cascallar (Eds.), *Optimising new modes of assessment: In search of qualities and standards* (Vol. 1, pp. 55–87): Springer Netherlands.

Topping, K. J. (2013). Peers as a source of formative and summative assessment. In J. H. McMillan (Ed.), *The SAGE handbook of research on classroom assessment (Vol* (pp. 395–412). Thousand Oaks, CA: Sage.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3*(4), 4–70.

Vanderhoven, E., Raes, A., Montrieux, H., Rotsaert, T., & Schellens, T. (2015). What if pupils can assess their peers anonymously? A quasi-experimental study. *Computers & Education, 81*, 123–32. doi:10.1016/j.compedu.2014.10.001.

Van Gennip, N. (2012). *Assessing together. Peer assessment from an interpersonal perspective.* (PhD), Universiteit Leiden.

van Gennip, N., Gijbels, D., Segers, M., & Tillema, H. H. (2010). Reactions to 360° feedback: The role of trust and trust-related variables. *International Journal of Human Resources Development and Management, 10*(4), 362–379. doi:10.1504/IJHRDM.2010.036088.

van Gennip, N., Segers, M., & Tillema, H. H. (2009). Peer assessment for learning from a social perspective: The influence of interpersonal variables and structural features. *Educational Research Review, 4*(1), 41–54. doi:10.1016/j.edurev.2008.11.002.

van Zundert, M., Sluijsmans, D., & van Merriënboer, J. (2010). Effective peer assessment processes: Research findings and future directions. *Learning and Instruction, 20*(4), 270–279. doi:10.1016/j.learninstruc.2009.08.004.

Yu, F. Y., & Wu, C. P. (2011). Different identity revelation modes in an online peer-assessment learning environment: Effects on perceptions toward assessors, classroom climate and learning activities. *Computers & Education, 57*(3), 2167–2177. doi:10.1016/j.compedu.2011.05.012.

Most relevant publications in the field of Psychology of Education:

By the first author (some):

Panadero, E., Alonso-Tapia, J., & Huertas, J. A. (2012). Rubrics and self-assessment scripts effects on self-regulation, learning and self-efficacy in secondary education. Learning and Individual Differences, 22(6), 806–813. doi: http://dx.doi.org/10.1016/j.lindif.2012.04.007

Panadero, E., & Jonsson, A. (2013). The use of scoring rubrics for formative assessment purposes revisited: A review. Educational Research Review, 9(0), 129–144. doi: 10.1016/j.edurev.2013.01.002

Panadero, E., Brown, G. T. L., & Courtney, M. G. R. (2014). Teachers' reasons for using self-assessment: A survey self-report of Spanish teachers. Assessment in Education: Principles, Policy & Practice, 21(3), 365–383. doi: 10.1080/0969594X.2014.919247

Panadero, E., & Järvelä, S. (2015). Socially shared regulation of learning: A review. European Psychologist, 20(3), 190–203. doi: 10.1027/1016-9040/a000226

Panadero, E. (2016). Is it safe? Social, interpersonal, and human effects of peer assessment: a review and future directions. In G. T. L. Brown & L. R. Harris (Eds.), *Human factors and social conditions of assessment*. New York: Routledge

By the second author (some):

Brown, G. T. L., Lake, R., & Matters, G. (2011). Queensland teachers' conceptions of assessment: The impact of policy priorities on teacher attitudes. Teaching and Teacher Education, 27(1), 210–220. doi: http://dx.doi.org/10.1016/j.tate.2010.08.003

Brown, G. T. L., & Harris, L. R. (2013). Student self-assessment. In J. McMillan (Ed.), The SAGE handbook of research on classroom assessment (pp. 367–393). Thousand Oaks, CA: SAGE.

Brown, G. T. L., Andrade, H., & Chen, F. (2015). Accuracy in student self-assessment: Directions and cautions for research. Assessment in Education: Principles, Policy & Practice, 1–14. doi: 10.1080/0969594X.2014.996523

Brown, G. T. L., Harris, L. R., O'Quin, C. R., & Lane, K. (2015). Using multi-group confirmatory factor analysis to evaluate cross-cultural research: identifying and understanding non-invariance. *International Journal of Research and Method in Education*. Advance online publication. doi: 10.1080/1743727X.2015.1070823.

Harris, L. R., & Brown, G. T. L. (2013). Opportunities and obstacles to consider when using peer- and self-assessment to improve student learning: case studies into teachers' implementation. *Teaching and Teacher Education, 36* 101–111. doi: 10.1016/j.tate.2013.07.008.

Harris, L. R., Brown, G. T. L., & Harnett, J. (2014). Analysis of New Zealand primary and secondary student peer- and self-assessment comments: applying Hattie & Timperley's feedback model. *Assessment in Education: Principles, Policy and Practice, 22*(2), 265–281. doi: 10.1080/0969594X.2014.976541.